WHAT IS CLAIMED IS:

1. A method of creating an audio-centric audio-visual summary of a video program, said video program having an audio track and an image track, said method comprising:

selecting a length of time L_{sum} of said audio-visual summary;

examining said audio track and image track;

identifying one or more audio segments from said audio track based on one or more predetermined audio, image, speech, and text characteristics which relate to desired content of said audio-visual summary, wherein said identifying is performed in accordance with a machine learning method which relies on previously-generated experience-based learning data to provide, for each of said audio segments in said video program, a probability that a given audio segment is suitable for inclusion in said audio-visual summary;

adding said audio segments to said audio-visual summary;

performing said identifying and adding in descending order of said probability until the length of time L_{sum} is reached; and

selecting only one or more image segments corresponding to the one or more identified audio segments, so as to yield a high degree of synchronization between said one or more audio segments and said one or more image segments.

2. A method as claimed in claim 1, wherein said identifying further comprises detecting audio segments comprising non-speech sounds; classifying said non-speech sounds according to contents; and, for each of said non-speech sounds, outputting a starting time code, length, and category.

3. A method as claimed in claim 2, wherein, when said audio segments comprise speech, said identifying comprises performing speech recognition on said audio segments to generate

speech transcripts, and outputting a starting time code and length for each of said speech transcripts.

4. A method as claimed in claim 3, wherein, when there is closed captioning present, said method further comprises aligning the closed captioning and the speech transcripts.

5. A method as claimed in claim 4, wherein said identifying further comprises generating speech units either based on said aligning, if said closed captioning is present, or based on said speech transcripts, if said closed captioning is not present, and creating a feature vector for each of said speech units.

6. A method as claimed in claim 5, further comprising computing an importance rank for each of said speech units.

7. A method as claimed in claim 6, further comprising receiving said speech units and determining identities of one or more speakers.

8. A method as claimed in claim 1, wherein said identifying further comprises segmenting said image track into individual image segments.

9. A method as claimed in claim 8, further comprising extracting image features and forming an image feature vector for each of said image segments.

10. A method as claimed in claim 9, further comprising determining identities of one or more faces for each of said image segments.

11. A method as claimed in claim 1, wherein said probability is computed in accordance with a method selected from the group consisting of a Naïve Bayes method, a decision tree method, a neural network method, and a maximum entropy method.

12. A method of creating an image-centric audio-visual summary of a video program, said video program having an audio track and an image track, said method comprising:

selecting a length of time L_{sum} of said audio-visual summary;

examining said image track and audio track of said video program;

identifying one or more image segments from said image track based on one or more predetermined image, audio, speech, and text characteristics which relate to desired content of said audio-visual summary, wherein said identifying is performed in accordance with a machine learning method which relies on previously-generated experience-based learning data to provide, for each of said image segments in said video program, a probability that a given image segment is suitable for inclusion in said audio-visual summary;

adding said one or more image segments to said audio-visual summary;

performing said identifying and adding in descending order of said probability until the length of time L_{sum} is reached; and

selecting only one or more audio segments corresponding to the one or more identified image segments, so as to yield a high degree of synchronization between said one or more image segments and said one or more audio segments.

13. A method as claimed in claim 12, wherein said identifying comprises segmenting said image track into individual image segments.

14. A method as claimed in claim 13, further comprising extracting image features and forming an image feature vector for each of said image segments.

15. A method as claimed in claim 14, further comprising determining identities of one or more faces for each of said image segments.

16. A method as claimed in claim 12, further comprising selecting a minimum playback time $L_{\min}$ for each of said image segments in said audio-visual summary.

17. A method as claimed in claim 16, wherein $L_{\min}$ is sufficiently small relative to L_{sum} such that a relatively large number of audio segments and image segments are provided in said audio-visual summary, to provide a breadth-oriented audio-visual summary.

18. A method as claimed in claim 16, wherein $L_{\min}$ is sufficiently large relative to L_{sum} such that a relatively small number of audio segments and image segments are provided in said audio-visual summary, to provide a depth-oriented audio-visual summary.

19. A method as claimed in claim 12, wherein said identifying further comprises detecting audio segments comprising non-speech sounds; classifying said non-speech sounds according to contents; and, for each of said non-speech sounds, outputting a starting time code, length, and category.

20. A method as claimed in claim 19, wherein, when said audio segments comprise speech, said identifying further comprises performing speech recognition on said audio segments to generate speech transcripts, and outputting a starting time code and length for each of said speech transcripts.

21. A method as claimed in claim 20, wherein, when there is closed captioning present, said method further comprises aligning the closed captioning and the speech transcripts.

22. A method as claimed in claim 21, wherein said identifying further comprises generating speech units either based on said aligning, if said closed captioning is present, or

based on said speech transcripts, if said closed captioning is not present, and creating a feature vector for each of said speech units.

23. A method as claimed in claim 22, further comprising computing an importance rank for each of said speech units.

24. A method as claimed in claim 23, further comprising receiving said speech units and determining identities of one or more speakers.

25. A method as claimed in claim 12, wherein said probability is computed in accordance with a method selected from the group consisting of a Naïve Bayes method, a decision tree method, a neural network method, and a maximum entropy method.

26. A method of creating an integrated audio-visual summary of a video program, said video program having an audio track and a video track, said method comprising:

selecting a length of time L_{sum} of said audio-visual summary;

selecting a minimum playback time $L_{\min}$ for each of said image segments to be included in the audio-visual summary;

creating an audio summary by selecting one or more desired audio segments until the audio-visual summary length L_{sum} is reached, said selecting being determined in accordance with a machine learning method which relies on previously-generated experience-based learning data to provide, for each of said audio segments in said video program, a probability that a given audio segment is suitable for inclusion in said audio-visual summary;

computing, for each of said image segments, a probability that a given image segment is suitable for inclusion in said audio-visual summary in accordance with said machine learning method;

for each of said audio segments that are selected, examining a corresponding image segment to see whether a resulting audio segment/image segment pair meets a predefined alignment requirement;

if the resulting audio segment/image segment pair meets the predefined alignment requirement, aligning the audio segment and the image segment in the pair from their respective beginnings for said minimum playback time L_{min} to define a first alignment point;

repeating said examining and aligning to identify all of said alignment points;

dividing said length of said audio-visual summary into a plurality of partitions, each of said partitions having a time period

either starting from a beginning of said audio-visual summary and ending at the first alignment point; or

starting from an end of the image segment at one alignment point, and ending at a next alignment point; or

starting from an end of the image segment at a last alignment point and ending at the end of said audio-visual summary; and

for each of said partitions, adding further image segments in accordance with the following:

identifying a set of image segments that fall into the time period of that partition;

determining a number of image segments that can be inserted into said partition;

determining a length of the identified image segments to be inserted;

selecting said number of the identified image segments in descending order of said probability that a given image segment is suitable for insertion in said audio-visual summary; and

from each of the selected image segments, collecting a section from its respective beginning for said time length and adding all the collected sections in ascending time order into said partition.

27. A method as claimed in claim 26, wherein said identifying further comprises detecting audio segments comprising non-speech sounds; classifying said non-speech sounds according to contents; and, for each of said non-speech sounds, outputting a starting time code, length, and category.

28. A method as claimed in claim 27, wherein, when said audio segments comprise speech, said identifying further comprises performing speech recognition on said audio segments to generate speech transcripts, and outputting a starting time code and length for each of said speech transcripts.

29. A method as claimed in claim 28, wherein, when there is closed captioning present, said method further comprises aligning the closed captioning and the speech transcripts.

30. A method as claimed in claim 29, further comprising generating speech units either based on said aligning, if said closed captioning is present, or based on said speech transcripts, if said closed captioning is not present, and creating a feature vector for each of said speech units.

31. A method as claimed in claim 30, further comprising computing an importance rank for each of said speech units.

32. A method as claimed in claim 31, further comprising receiving said speech units and determining identities of one or more speakers.

33. A method as claimed in claim 26, wherein $L_{\min}$ is sufficiently small relative to L_{sum} such that a relatively large number of image segments are provided in said audio-visual summary, to provide a breadth-oriented audio-visual summary.

34. A method as claimed in claim 26, wherein $L_{\min}$ is sufficiently large relative to L_{sum} such that a relatively small number of image segments are provided in said audio-visual summary, to provide a depth-oriented audio-visual summary.

35. A method as claimed in claim 26, wherein said probability that said given audio segment is suitable for inclusion in said audio-visual summary is computed in accordance with a method selected from the group consisting of a Naïve Bayes method, a decision tree method, a neural network method, and a maximum entropy method.

36. A method as claimed in claim 26, wherein said probability that said given image segment is suitable for inclusion in said audio-visual summary is computed in accordance with a method selected from the group consisting of a Naïve Bayes method, a decision tree method, a neural network method, and a maximum entropy method.

37. A method as claimed in claim 26, wherein said identifying further comprises segmenting said image track into individual image segments.

38. A method as claimed in claim 37, further comprising extracting image features and forming an image feature vector for each of said image segments.

39. A method as claimed in claim 38, further comprising determining identities of one or more faces for each of said image segments.

40. A method of creating an audio-centric audio-visual summary of a video program, said video program having an audio track and an image track, said method comprising:

selecting a length of time L_{sum} of said audio-visual summary;

examining said audio track and image track;

identifying one or more audio segments from said audio track based on one or more predetermined audio, image, speech, and text characteristics which relate to desired content of said audio-visual summary, wherein said identifying is performed in accordance with a predetermined set of heuristic rules to provide, for each of said audio segments in said video program, a ranking so as to determine whether a given audio segment is suitable for inclusion in said audio-visual summary;

adding said audio segments to said audio-visual summary;

performing said identifying and adding in descending order of said ranking of audio segments until the length of time L_{sum} is reached; and

selecting only one or more image segments corresponding to the one or more identified audio segments, so as to yield a high degree of synchronization between said one or more audio segments and said one or more image segments.

41. A method as claimed in claim 40, wherein said identifying further comprises detecting audio segments comprising non-speech sounds; classifying said non-speech sounds according to contents; and, for each of said non-speech sounds, outputting a starting time code, length, and category.

42. A method as claimed in claim 41, wherein, when said audio segments comprise speech, said identifying comprises performing speech recognition on said audio segments to generate speech transcripts, and outputting a starting time code and length for each of said speech transcripts.

43. A method as claimed in claim 42, wherein, when there is closed captioning present, said method further comprises aligning the closed captioning and the speech transcripts.

44. A method as claimed in claim 43, further comprising generating speech units either based on said aligning, if said closed captioning is present, or based on said speech transcripts, if said closed captioning is not present, and creating a feature vector for each of said speech units.

45. A method as claimed in claim 44, further comprising receiving said speech units and determining identities of one or more speakers.

46. A method as claimed in claim 40, wherein said identifying comprises segmenting said image track into individual image segments.

47. A method as claimed in claim 46, further comprising extracting image features and forming an image feature vector for each of said image segments.

48. A method as claimed in claim 47, further comprising determining identities of one or more faces for each of said image segments.

49. A method as claimed in claim 40, further comprising computing said ranking for each of said speech units.

50. A method of creating an image-centric audio-visual summary of a video program, said video program having an audio track and an image track, said method comprising:

selecting a length of time L_{sum} of said summary;

examining said image track and audio track;

identifying one or more image segments from said image track based on one or more predetermined image, audio, speech, and text characteristics which relate to desired content of said audio-visual summary, wherein said identifying is performed in accordance with a predetermined set of heuristic rules to provide, for each of said image segments in said video program, a ranking so as to determine whether a given image segment is suitable for inclusion in said audio-visual summary;

adding said one or more image segments to said audio-visual summary;

performing said identifying and adding in descending order of said ranking until the length of time L_{sum} is reached; and

selecting only one or more audio segments corresponding to the one or more identified image segments, so as to yield a high degree of synchronization between said one or more image segments and said one or more audio segments.

51. A method as claimed in claim 50, wherein said identifying comprises clustering image segments of said video program based on predetermined visual similarity and dynamic characteristics.

52. A method as claimed in claim 51, wherein said identifying comprises segmenting said image track into individual image segments.

53. A method as claimed in claim 52, further comprising extracting image features and forming an image feature vector for each of said frame clusters.

54. A method as claimed in claim 53, further comprising determining identities of one or more faces for each of said frame clusters.

55. A method as claimed in claim 50, wherein said identifying further comprises detecting audio segments comprising non-speech sounds, classifying said non-speech sounds according to contents; and, for each of said non-speech sounds, outputting a starting time code, length, and category.

56. A method as claimed in claim 55, wherein, when said audio segments comprise speech, said identifying comprises performing speech recognition on said audio segments to generate speech transcripts, and outputting a starting time code and length for each of said speech transcripts.

57. A method as claimed in claim 56, wherein, when there is closed captioning present, said method further comprises aligning the closed captioning and the speech transcripts.

58. A method as claimed in claim 57, further comprising generating speech units either based on said aligning, if said closed captioning is present, or based on said speech transcripts, if said closed captioning is not present, and creating a feature vector for each of said speech units.

59. A method as claimed in claim 58, further comprising computing an importance rank for each of said speech units.

60. A method as claimed in claim 59, further comprising receiving said speech units and determining identities of one or more speakers.

61. A method as claimed in claim 50, further comprising selecting a minimum playback time $L_{\min}$ for each of said image segments in said audio-visual summary.

62. A method as claimed in claim 61, wherein $L_{\min}$ is sufficiently small relative to L_{sum} such that a relatively large number of audio segments and image segments are provided in said audio-visual summary, to provide a breadth-oriented audio-visual summary.

63. A method as claimed in claim 61, wherein $L_{\min}$ is sufficiently large relative to L_{sum} such that a relatively small number of audio segments and image segments are provided in said audio-visual summary, to provide a depth-oriented audio-visual summary.

64. A method of creating an integrated audio-visual summary of a video program, said video program having an audio track and a video track, said method comprising:

selecting a length L_{sum} of said audio-visual summary;

selecting a minimum playback time $L_{\min}$ for each of a plurality of image segments to be included in the audio-visual summary;

creating an audio summary by selecting one or more desired audio segments, said selecting being determined in accordance with a predetermined set of heuristic rules to provide, for each of said audio segments in said video program, a ranking to determine whether a given audio segment is suitable for inclusion in said video summary;

performing said selecting in descending order of said ranking of audio segments until said audio-visual summary length is reached;

grouping said image segments of said video program into a plurality of frame clusters based on a visual similarity and a dynamic level of said image segments, wherein each frame

cluster comprises at least one of said image segments, with all the image segments within a given frame cluster being visually similar to one another;

for each of said audio segments that are selected, examining a corresponding image segment to see whether a resulting audio segment/image segment pair meets a predefined alignment requirement;

if the resulting audio segment/image segment pair meets the predefined alignment requirement, aligning the audio segment and the image segment in the pair from their respective beginnings for said minimum playback time L_{min} to define a first alignment point;

repeating said examining and aligning to identify all of said alignment points;

dividing said length of said audio-visual summary into a plurality of partitions, each of said partitions having a time period

either starting from a beginning of said audio-visual summary and ending at the first alignment point; or

starting from an end of the image segment at one alignment point, and ending at a next alignment point; or

starting from an end of the image segment at a last alignment point and ending at the end of said audio-visual summary; and

dividing each of said partitions into a plurality of time slots, each of said time slots having a length equal to said minimum playback time L_{min};

assigning said frame clusters to fill said time slots of each of said partitions based on the following:

assigning each frame cluster to only one time slot; and

maintaining a time order of all image segments in the audio-visual summary;

wherein said assigning said frame clusters to fill said time slots is performed in accordance with a best matching between said frame clusters and said time slots.

65. A method as claimed in claim 64, wherein said best matching is computed by a method of maximum-bipartite-matching.

66. A method as claimed in claim 65, wherein, if there are more time slots than frame clusters, identifying those frame clusters which contain more than one image segment, and assigning image segments from said identified frame clusters to time slots until all of said time slots are filled, while maintaining said time order of said image segments in said audio-visual summary.

67. A method as claimed in claim 66, further comprising reviewing said audio-visual summary to ensure that said time order is maintained, and, if said time order is not maintained, reordering said image segments that were added in each partition so that said time order is maintained.

68. A method as claimed in claim 64, wherein said identifying further comprises detecting audio segments comprising non-speech sounds, classifying said non-speech sounds according to contents; and, for each of said non-speech sounds, outputting a starting time code, length, and category.

69. A method as claimed in claim 68, wherein, when said audio segments comprise speech, said identifying comprises performing speech recognition on said audio segments to generate speech transcripts, and outputting a starting time code and length for each of said speech transcripts.

70. A method as claimed in claim 69, wherein, when there is closed captioning present, said method further comprises aligning the closed captioning and the speech transcripts.

71. A method as claimed in claim 70, further comprising generating speech units either based on said aligning, if said closed captioning is present, or based on said speech transcripts, if said closed captioning is not present, and creating a feature vector for each of said speech units.

72. A method as claimed in claim 71, further comprising computing an importance rank for each of said speech units.

73. A method as claimed in claim 72, further comprising receiving said speech units and determining identities of one or more speakers.

74. A method as claimed in claim 64, wherein $L_{\min}$ is sufficiently small relative to L_{sum} such that a relatively large number of image segments are provided in said audio-visual summary, to provide a breadth-oriented audio-visual summary.

75. A method as claimed in claim 64, wherein $L_{\min}$ is sufficiently large relative to L_{sum} such that a relatively small number of image segments are provided in said audio-visual summary, to provide a depth-oriented audio-visual summary.

76. A method as claimed in claim 64, wherein said identifying comprises segmenting said image track into individual image segments.

77. A method as claimed in claim 76, further comprising extracting image features and forming an image feature vector for each of said frame clusters.

78. A method as claimed in claim 77, further comprising determining identities of one or more faces for each of said image segments.